MANAV RACHNA INTERNATIONAL UNIVERSITY

Short Synopsis

For

Ph. D. Programme 2013-14

Title: Design and Development of Efficient Clustering Techniques in Data Mining

DEPARTMENT OF COMPUTER SCIENCE & ENGINEERING

FACULTY OF ENGINEERING & TECHNOLOGY

Submitted by:

Name: Anupama Luthra
Registration No.: 13/Ph.D./015

Supervisor :

Name: Dr. Suresh Kumar
Designation: Professor

ABSTRACT

K-Means is a widely used partition based clustering algorithm which organizes input dataset into predefined number of clusters. Simplicity and speed in classification of massive data are two features which have made K-Means a very popular algorithm. The original K-Means algorithm clusters numerical data, however its extensions K-Modes and K-Prototype work on categorical and mixed data sets respectively. K-Means has a major limitation -- the number of clusters, K, need to be pre-specified as an input to the algorithm. In absence of thorough domain knowledge, or for a new and unknown dataset, this advance estimation and specification of cluster number typically leads to forced clustering of data, and proper classification does not emerge. Also one new need with K-Prototype is to find a unified metric for both numerical and categorical data.

The author will propose algorithms based on the K-Means and its extensions K-Modes and K-Prototype, but having advance features of intelligent data analysis and automatic generation of appropriate number of clusters. The automatic and intelligent cluster generation by the proposed algorithm will be compared against the results obtained with ideally pre-optimized number of clusters specified in K-Means .This will be done using many different real data sets. Also a unified similarity metric that works with mixed data sets will be proposed.

Keywords: *Clustering, K-Means, K-Modes, K-Prototype, Dependency, Prior input, Number of clusters, Unified similarity metric*

CONTENTS

S.No.	Description	Page No.
1	Introduction	1-2
2	Literature Review	3-5
	2.1 Extension of K-Means	3-4
	2.2 Extension of K-Modes and K-Prototype	4-5
	2.3 Unified similarity metric for mixed data sets	5
3	Description of Broad Area	6-15
	3.1 Introduction to Data Mining	6-7
	3.2 Data Mining Techniques	7-9
	3.2.1 Association Rule Analysis	7-8
	3.2.2 Classification	8-9
	3.2.3 Clustering	9
	3.3 Clustering Methods	9-11
	3.3.1 Partitioning Method	9-10
	3.3.2 Hierarchical Method	10
	3.3.3 Density Based Method	10-11
	3.3.4 Grid-Based Method	11
	3.4 Introduction to K-Means algorithm	11-12
	3.5 K-Modes algorithm	13
	3.6 K-Prototype algorithm	14-15
4	Objectives of the Study	15
5	Methodology to be adopted	16
6	Expected outcome of the research	16
7	References	17-18

1. INTRODUCTION

In today s globalized and increasingly smaller world, the markets are barrierless and the reach of businesses is expanding beyond cities to entire nations, and even across the globe. Large organizations in Banking, Automobile, Agriculture, Education etc. use huge amount of data and classify it to understand demographies, consumption, usage patterns etc. These businesses also generate huge amount of data themselves, which is increasing exponentially. For taking right decisions at right time, this huge amount of data needs to be processed efficiently and accurately for correct interpretations and decisions. This entire philosophy of storing, maintaining, classifying & interpreting data, to find patterns or trends for better business decisions, is an upcoming area of research.

To help organizations take right decisions at right time, Data mining is a tool which provides techniques to process large amount of data efficiently and presents it in the required form. Data Mining is a process of drawing out useful patterns or knowledge from the huge data collected in information systems and to use these patterns in taking safe and smart decisions. The predefined methods and algorithms that are used to extract these useful patterns are together called Data Mining Techniques. Some popular data mining techniques include Frequent Pattern Mining, Association Rule Analysis, Classification and Clustering.

Clustering is a technique of segregating the objects into partitions such that the objects in a one group are more similar to each other than the objects in the other group. Clustering has its applications in variety of domains like psychology, statistics, medicine, engineering, computer science etc. For example, in an organization, grouping and identifying the products which are not in high demand may help in reducing their production to cut losses. Further, in Educational Institutions, grouping the students according to their academic performance may help in identifying the students with lower grades. These students can be motivated to attend remedial classes to overcome their difficulties. In Health sector, data mining and clustering may help in identifying the link between the disease symptoms. In Banking sector, it can be used to group customers with overdue credit card payments. In Market Research, clustering can be used to identify customers having certain buying patterns. A lot of work is being done to apply this technique in various other areas, apart from the above mentioned.

Many clustering algorithms have been proposed in the literature [7,20]. These clustering algorithms are broadly classified into two categories: Hierarchical and Partitional. The hierarchical algorithms find clusters by arranging them into hierarchy (top-down or bottom-up), as a result they are not suitable for large data sets. On the other hand, the partition based clustering algorithms find the clusters independently. So they can easily partition large datasets. In this method, the given dataset of n objects is partitioned into k groups or clusters where $k \leq n$ with the constraint that each group must be comprised of at least one object and each object is a member of only one cluster. Partition based method is an iterative method in which the clusters once created are further improved by shifting the objects from one cluster to another depending upon value of some objective function. K-means algorithm is one of the commonly used techniques in this category.

K-Means is a simple algorithm known for its speed. The algorithm is not expensive in terms of cost and works well with high dimensional and large data sets. However there exist some limitations in this algorithm. One major limitation is that the clusters produced are highly dependent on the objects initially selected as centroids (cluster centers). As the initial centroids are selected randomly, K-Means algorithm may not provide same result for different runs on same data set. A lot of work has been done to overcome this limitation.

Another limitation is the requirement to specify a pre-defined value of K (number of clusters) as input. This is domain specific, and if the person using the algorithm is not domain expert, then an incorrect number of clusters may be input, leading to inefficient grouping of data. To overcome this limitation, researchers are still exploring ways.

To overcome the limitations of K-mean algorithm, the author will propose algorithms based on K-Means and its extensions K-Modes and K-Prototype algorithm for numerical, categorical and mixed data sets which does not require the value of K as input. In order to increase the accuracy of the clusters produced by the K-Prototype algorithm a similarity metric that works with mixed data sets will be proposed.

2. LITERATURE REVIEW

The Literature Review has been divided into three sections. Section 2.1 deals with the attempts that have been made in the literature to remove the limitation of giving the number of clusters required as an input for numerical data. Section 2.2 discusses the work done in the field of removing the limitation of providing the number of clusters required (K) in the K-Means for categorical and mixed data. Section 2.3 discusses the similarity metrics that have been suggested for mixed data sets.

2.1 Contribution of some authors to remove the limitation of providing the value of K initially for numerical data is discussed below:

Pelleg Dan et al. [16] suggested X-Means algorithm as an extension of K-Means which required the user to input a range representing the lower and upper value of K instead of a particular value of K. The algorithm initially takes lower bound of the given range as K and continues to add centroids until the upper bound is reached. The algorithm terminates as soon as it gets the centroid set that scores the best. The drawback lies in the fact that it requires the user to input a range suggesting the lower and upper bound of K.

Tibshirani R. et al. [18] used the technique of Gap Statistic. In this technique the quality of clusters produced is verified using an appropriate reference distribution. The algorithm works well with separated clusters.

Wagstaff Kiris et al. [19] suggested to utilize information about the problem domain in order to put some constraints on the data set. During the clustering process it is ensured that none of the constraint is violated. This algorithm requires some domain specific information, which sometimes becomes difficult to obtain.

Cheung Yiu-Ming [5] proposed an extension of K-Means clustering technique named STep-wise Automatic Rival penalized (STAR) K-Means algorithm to overcome the major limitations of K-Means. In the first step of the algorithm, cluster centers are provided and in

the second step the units are adjusted adaptively by a learning rule. The limitation of this algorithm is the complex computation involved in it.

Shafeeq Ahamed B.M. et al. [3] proposed an algorithm in which the appropriate number of clusters is found dynamically. The main drawback of this approach is that its computational time is more than the K-Means for larger data sets. Also the user has to input the value of K as 2 in the first run.

Leela V.et al. [12] proposed Y-means algorithm. Initially, clusters were found using K-Means algorithm on the data set. A sequence of splitting, deleting and merging the clusters was then followed to find the optimal number of clusters. The limitation of this algorithm is that it depends on K-Means algorithm to find the initial clusters.

Abubaker Mohamed et al. [1] presented an approach based on the K-Nearest Neighbor method. The only input parameter taken by the algorithm was k_n (the number of nearest neighbor). The drawback of this algorithm is that it require the number of nearest neighbors k_n as input.

2.2 This section discusses the work that has been done to overcome the limitation of providing the number of cluster required (K) in the K-Means for categorical and mixed data sets .

San Mar Ohn et al. [17] proposed an algorithm in which a parameter was used to control the number of clusters in the clustering process. A suitable value of regularization parameter was chosen to find the most stable clustering results. The major limitation of the proposed algorithm is that an input parameter representing the initial cluster centers is required.

Cheung Yiu-ming et al.[6] presented penalized competitive learning algorithm that require some initial value of K which should not be less than the original value of K. The resulting clusters are more accurate than the original K-Modes and K-Modes with Ng s dissimilarity metric proposed by H. Liao and M.K. Ng [14]. But this algorithm has much computation involved in it.

Liang Jiye et al. [13] extended K-Prototype algorithm by proposing a new dissimilarity measure for mixed data set. The measures of within-cluster entropy and between-cluster entropy were used to identify the clusters with minimum coherence in a mixed dataset. The major limitation of this algorithm is that it requires input parameters representing the minimum and maximum number of clusters that can be generated from the data set.

Ahmad Amir et al. [4] proposed a new cost function for mixed data set. The authors extended K-Means algorithm that worked well with mixed datasets by introducing a new distance measure and a new way of finding the centroids of the clusters. As the clustering process is based on the significance of an attribute in the dataset so the computation times of the algorithm increases with the increase in the dimensions of the data.

2.3 Attempts have been done and still going on to find a similarity metric that works with data sets containing mixed attributes. Some of the work is discussed below:

Ahmad Amir et al. [4] proposed a similarity measure for mixed datasets. The distance measure for categorical attributes was based on the distance between every two attribute values of every attribute. The numerical attributes were normalized and discretized to find added weight (w_t) to be added in their distance measure. Since in this method distance is calculated for each attribute, it is not suitable for noisy and high dimensional datasets. The results obtained by this algorithm can be improved further by improving discretizing methods for numeric valued attributes.

Cheung Yiu-ming et al. [6] proposed a unified metric for mixed data sets. The author treated the categorical and numerical attributes differently. The categorical attributes were considered one by one, while all the numerical attributes were handled together as a vector in the clustering process. The use of this unified metric increases computation time.

3. DESCRIPTION OF BROAD AREA

3.1 Data Mining

In today s world information technology has affected every aspect of life be it be food preparing gadgets in mother s kitchen, maintaining financial records in banks, maintaining health records of the patients in hospitals, recording academic performance of the students in educational institutes. In late 1980s a new trend has come up, identifying the meaningful data collected in information systems, finding useful patterns to take smart and safe decisions based on these patterns. The process of extracting useful patterns or knowledge is called Data Mining or KDD (Knowledge Discovery from Data). In Data mining, the process of knowledge discovery is shown in Figure 1.

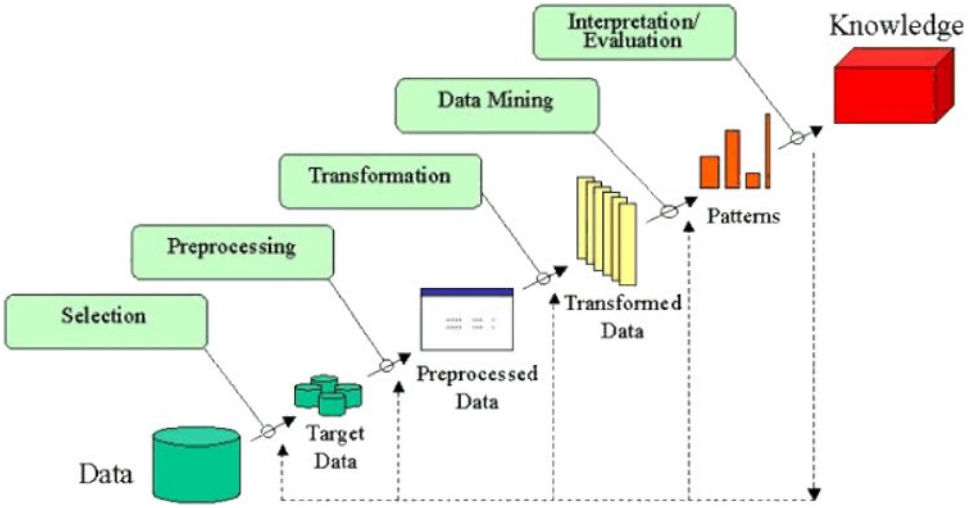

Figure 1: Data Mining as a step in Knowledge Discovery [8]

Simply, data mining can be defined as the process of automatic detection of relevant patterns in a database consisting of current and historical data, using predefined approaches and algorithms (together called data mining techniques).Data mining uses approaches and methods from multi-disciplinary areas out of which Statistics and Artificial Intelligence (AI) are the major ones. Statisticians have used sophisticated techniques to analyze data and provide business projections. So Data Mining can be considered as computer-assisted method to explore the data statistically. Also some of the techniques used in data mining are borrowed from the field of AI. The major difference between the data mining techniques is

that data mining techniques deal with huge stores of data while AI techniques deal with those data sets that fit in the main memory of the computer.

This development of data warehousing and data mining techniques has enabled the organization to change the pattern of their queries. Earlier, the queries of the companies were of the form "What was the total number of sales in City A in the month of April?" Now they have the answers to queries such as: "Which product should be launched in the coming year to increase the business of the company? . Most of the business sectors are making use of data mining techniques to make analytical decisions be it a Health sector, Agriculture, Banking sector, Educational organizations or any other sector requiring an analysis into the current and historical data.

3.2 Data Mining Techniques

The next section gives brief overview of the various popular data mining techniques.

3.2.1 Frequent Pattern Mining and Association Rule Analysis

In this technique the patterns that appear in the data set frequently are extracted and association between these frequent patterns is find out by generating Association Rules. The most popular example of this technique is Market Basket Analysis. In this example, the items that are frequently purchased by the customers are find out from the data set of their buying habits, the association rules are then generated using these itemsets which further elaborate the relation between these frequent item sets.

The Association Rule is of the form:

buys(X, bread) => buys(X, jam) [support=20%, confidence=60%]

This rule shows that a person who buys bread tends to buy jam at the same time. The association rules are measured by their interestingness parameter, which in turn is represented by two measures Support and Confidence. These measures are expressed in % as shown above in the given example. For the above association rule a support of 20% reflects that 20% of the customers buy bread and jam together, a confidence of 60% shows that 60% of the customers who have bought bread are likely to buy jam . From this association rule, the manager of the store can reorganize the things by putting bread and

jam together at the same place so that the persons buying the bread are prompted to buy the jam also, this in turn will increase the sale of jam . The Apriori algorithm is one of commonly used technique for finding frequent item sets from the given data set, from these item sets Association rules can be generated.

3.2.2 Classification

As the name suggests, this techniques is used to divide or classify the data into different groups. Given a dataset with at least one of the attributes as categorical attribute which will represent the class, this technique will divide the dataset into groups where number of groups depends on the number of values in the categorical attribute. This process of classification in groups is done in two steps:

i) Generate Classification Rules

ii) Classify the data

The attributes in the dataset used for generating Classification rules are called Input parameters and the attribute which represents the class is called Target attribute.

This can be better understood with the following example:

Table 1: Dataset for Classification

Age	Income	Applies_for_Loan
Senior	Low	No
Middle_aged	Low	Yes
Senior	High	Yes
Middle_aged	High	Yes

Here Age and Income are Input parameters and Applies_for_Loan is Target parameter.
In the first step, two Classification Rules are generated:

If Age= Senior and Income=High Then Applies_for_Loan=Yes

If Age=Middle_Aged Then Apples_for_Loan=Yes

In the second step, the classification of the data whose class is unknown is performed. Suppose we add a new object in the dataset with value of **Applies_for_Loan** as unknown

then based on the above generated classification rules the missing value will be predicted as shown below:

Table 2: Dataset for predicting the class

Age	Income	Applies_for_Loan
Senior	Low	No
Middle_aged	Low	Yes
Senior	High	Yes
Middle_aged	High	Yes
Senior	High	?

According to the generated classification rules, the missing value will be predicted as **Yes .** There exist various ways in which the classification can be performed. Some of the popular algorithms are ID3, C4.5, Fuzzy C4.5.

3.2.3 Clustering

Clustering is a technique of segregating the objects into partitions such that the objects in a certain cluster are more similar to each other than the objects in the other clusters. It differs from Classification in a way that it is unsupervised while Classification is supervised. The following section discusses the clustering methods.

3.3 Clustering Methods

Number of methods have emerged to partition the data and have been successfully applied to real life data mining problems. These methods are discussed below:

3.3.1 Partitioning Method

In this method, the given dataset of n objects is partitioned into k groups or clusters where $k \leq n$ with the constraint that each group must be comprised of at least one object and each object is a member of only one cluster. This is an iterative approach in which the clusters are refined in every step by making the objects move from one cluster to another depending upon value of some objective function. The famous algorithms that fall in this category are: K-Means and K-Medoids.

Limitation and Advantage:

The major limitation of this approach is that the algorithms under this category work well when clusters are globular. Also this method is sensitive to noise. The advantage is that all the algorithms falling in this category are scalable.

3.3.2 Hierarchical Method

In this method the objects in the given data set are organized into a tree of clusters. There are two ways of using this method :

Agglomerative method: In this method every object is put into a separate cluster and then a process of merging these unit clusters into large clusters is initiated until a single cluster containing all the objects is formed. AGNES (Agglomerative Nesting) algorithm falls under this category.

Divisive method: This is a reverse of Agglomerative method. This method starts with all the objects in a single cluster and then dividing into smaller clusters until the unit clusters are obtained. DIANA (Divisive Analysis) comes under this category.

Limitation and Advantage:

The limitation of this method is that this approach works well when the clusters are globular. The algorithms under this category are sensitive also to noise and outliers. The advantage of this method is that any number of clusters desired by the user can be obtained by chopping the tree of clusters.

3.3.3 Density Based Method

This method requires an initial input of minimum number of objects needed to make a cluster and the minimum distance between objects within a cluster. This method is helpful in removing the outliers as the outliers will themselves form a dense cluster. The famous algorithms of this category are: DBSCAN and OPTICS

Limitation and Advantage:

The limitation of the algorithms under this category are they require some input parameters which sometimes becomes difficult to predict. The advantage of this approach is that it helps in detecting outliers by making a separate cluster of outliers. Also this method works well with clusters of arbitrary shapes.

3.3.4 Grid-Based Method

As suggested by name of this method, it uses a multi resolution grid structure. The objects in the dataset are divided into a multilevel hierarchical grid structure. The clusters produced by this method are isothetic, that is, all of the cluster boundaries are either vertical or horizontal and no diagonal boundary. This constraint can be removed by combining the approaches of Density-Based and Grid-Based method as done in WaveCluster algorithm.

Limitation and Advantage:

The limitation of this method is that the clusters produced are isothetic and hence of low quality. The major advantage is that this method is efficient in terms of time complexity. The famous algorithm of this category is: STING.

The author will be focusing on the famous partitioning method based clustering algorithm: K-Means and its extensions, K-Modes and K-Prototype.

3.4 Introduction to K-Means algorithm

The K-Means is a partition based iterative algorithm using Euclidean distance as an objective function. It groups a given data set into certain number of clusters K (K in K-Means represents number of clusters required). The procedure starts with the selection of K centroids for K clusters. The clusters and their centroids are recomputed until all the data points in each cluster are at their minimum distance from their centroids. In this technique Euclidean distance is taken as given in Equation 1.

$$d_{euc}(x, y) = \sum_{i=i}^{n} \sqrt{(x_i - y_i)^2} \qquad (1)$$

The basic algorithm works as follows: [2]

Algorithm 1 : Basic K-Means **Input:** Numerical **d**ataset (D), Number of clusters (K) **Output:** Elements of dataset classified into K clusters
1. Select random K points as initial Centroids 2.Repeat 3.Create K clusters by assigning all data points to the closest centroid 4.Recompute centroids for each cluster to improve accuracy 5.Until the cluster centroids don t change

The K-Means algorithm is chosen by the author because of the following reasons:

- The K-Means algorithm is not very expensive in terms of time.
- Works well with high dimensional data and large data sets.
- Produces highly cohesive clusters.

Apart from these benefits it suffers from these limitations:

- The initial centroids are chosen randomly so the final clusters produced are highly sensitive to initial centroids.
- As the centroids are the mean of the values lying in that cluster, so sensitive to outliers and produces spherical clusters only.
- Value of K is required to be input, which sometimes becomes domain specific and if the person using the algorithm is not domain expert then it can cause problems.

3.5 K-Modes algorithm

K-Means has gain popularity because of its simplicity and speed of classifying massive data rapidly and efficiently. The K-Modes extends the K-Means algorithm to cluster categorical data with the following approach [11]:

- A Simple matching dissimilarity function suitable for categorical data is used instead of Euclidean distance.
- Modes are used to represent centroids instead of Mean values.
- A frequency based method is used to find centroids in each iteration of the algorithm.

Basic K-Modes works as follows: [9]

Algorithm : Basic K-Modes

Input: Categorical dataset (D), Number of clusters (K)

Output: Elements of dataset classified into K clusters

1. Select *k* initial modes, one for each cluster.
2. Allocate an object to the cluster whose mode is the nearest to it according to:

$$?_?(?, ?) = \sum_{??}^{?} ?(?_?, ?_?)$$

$?(?_?, ?_?)=0$ if $x_j=y_j$ else 1

Update the mode of the cluster after each allocation.

3. After all objects have been allocated to clusters, retest the dissimilarity of objects against the current modes. If an object is found such that its nearest mode belongs to another cluster rather than its current one, reallocate the object to that cluster and update the modes of both clusters.
4. Repeat 3 until no object has changed clusters after a full cycle test of the whole data set.

Since K-Modes follows the same iterative process as followed by K-Means to produce clusters so it carries in itself the merits and demerits of K-Means.

3.6 K-Prototype algorithm

Earlier, clustering techniques were developed focusing on a single type of attributes, either numerical or categorical. Since mixed data set is common in real life, so techniques need to be developed to group this type of data. The techniques used only for numerical data or for categorical data cannot be directly applied as they differ in their behavior. The numerical data is continuous whereas categorical attributes values are not only discontinuous but also disordered.

K-Prototype is a variant of K-Means that can be used with numeric or categorical data sets. K-Prototype extends the idea of K-Means by applying Euclidean distance to numeric attributes and Binary distance to categorical attributes. Basic K-Prototype algorithm works as follows: [10]

Algorithm : Basic K-Prototype

Input: Mixed dataset (D), Number of Clusters (K)

Output: Elements of Dataset classified into K clusters

1. Select *k* initial prototypes from a data set D, one for each cluster.

2. Allocate each object in D to a cluster whose prototype is the nearest to it according to the following distance measure:

$$?(?_?, ?_?) = ?_{??} (?_{??}^{?} - ?_{??}^{?}) + ?_? ?_{??} ?(?_{??}^{?}, ?_{??}^{?})$$

where $\delta(p,q)=0$ for $p=q$ and $\delta(p,q)=1$ for $p \neq q$.

$?_{??}^{?}$ and $?_{??}^{?}$ are values of numeric attributes, whereas $?_{??}^{?}$ and $?_{??}^{?}$ are values of categorical attributes for object I and the prototype of cluster *l*.

m_r and m_c are the numbers of numeric and categorical attributes.

γ_l is a weight for categorical attributes for cluster *l*.

Update the prototype of the cluster after each allocation.

3. After all objects have been allocated to a cluster, retest the similarity of objects against the current prototypes. If an object is found such that its nearest prototype belongs to

another cluster rather than its current one, reallocate the object to that cluster and update the prototypes of both clusters.

4. Repeat (3) until no object has changed clusters after a full cycle test of X.

However, the Binary distance for categorical attributes does not represent the real situation as the categorical values may have some other degree of difference rather than just 0 or 1. So, various extensions of the K-Prototype have been proposed. All these extensions suffer from the limitation of inputting the number of clusters required. Since K-Prototype extends the ideas of K-Means, the same weaknesses of K-Means are retained. The focus of research work will be on removal of one of the common limitation beard by the above discussed three algorithms of inputting the required number of clusters (K) to the algorithms.

4. OBJECTIVES OF OUR STUDY

To design an algorithm based on the K-Means for numerical data sets to overcome the limitation of providing the number of clusters at the very beginning based on anticipation.

To modify the K-Modes algorithm to cluster categorical data sets so as to overcome the limitation of inputting the number of clusters required initially.

To transform the K-Prototype algorithm for mixed data sets to overcome the limitation of providing the number of clusters required initially.

The compare the accuracy of the clusters produced by the proposed algorithms with that of the original K-Means, K-Modes and K-Prototype algorithm on different real world datasets of different sizes and dimensions from UCI machine Repository using RapidMiner.

To develop a suitable unified similarity metric for mixed data sets.

To use the proposed unified similarity metric in K-Prototype algorithm.

To modify the K-Prototype algorithm to overcome the limitation of providing the number of clusters initially by using this unified similarity metric.

5. METHODOLOGY

- Algorithms will be proposed based on the K-Means and its extensions K-Modes and K-Prototype for numerical, categorical and mixed data sets which divides the input data set into appropriate clusters without taking number of clusters K as input, as it was required in the original algorithms.
- A similarity metric that works with mixed data sets will be proposed and that metric will be used in basic K-Prototype and proposed K-Prototype algorithms.
- The accuracy of the clusters produced will be compared with that of the original algorithms. To achieve this we will be implementing the modified algorithms in C# and will compare the accuracy of the clusters with original algorithms using RapidMiner. For this purpose many real data sets will be used from UCI Machine Repository (A website that maintains 300 data sets as a service to the machine learning community).

6. EXPECTED OUTCOME OF THE RESEARCH

New algorithms based on the K-Means, K-Modes and K-Prototype but with advance features of intelligent data analysis and automatic generation of appropriate number of clusters for numerical, categorical and mixed data sets. The proposed algorithms will also generate clusters that are more accurate as compared to the clusters generated by the original algorithms. Our research work will also come up with a new unified similarity metric that works for mixed data sets in data clustering. This metric will be implemented in K-Prototype algorithm and the accuracy of the clusters will be improved.

REFERENCES

1. Abubaker, Mohamed, Ashour, Wesam (2013). Efficient Data Clustering Algorithms: Improvements over K means. International Journal of Intelligent Systems and Applications, 5(3), 37-49.

2. Agha, El, Mohammed, Ashour, M., Wesam (2012). Efficient and Fast Initializtion Algorithm for K-means Clustering. I.J. Intelligent Systems and Applications. 4(1), 21-31.

3. Ahamed, Shafeeq, B., M., Hareesha, K., S.(2012). Dynamic Clustering of Data with Modified K-Means Algorithm. International Conference on Information and Computer Networks (ICICN 2012). 27, 221-225.

4. Ahmad, A., Dey, L. (2007). A K-Mean Clustering Algorithm for Mixed Numeric and Categorical Data. Data & Knowledge Engineering. 63, 503 527.

5. Cheung, Yiu-Ming (2013). k*-Means: A new generalized k-means clustering algorithm. Pattern Recognition Letters. 24, 2883 2893.

6. Cheung, Yiu-ming, Jia, Hong (2013). Categorical-and-numerical-attribute data clustering based on a unified similarity metric without knowing cluster number. Pattern Recognition, 46, 2228 2238.

7. Gan, Guojun, Ma, Chaoqun, Wu, Jianhong (2007).Data clustering: theory, algorithms, and applications. SIAM: Society for Industrial and Applied Mathematics.

8. Han, Jiawei, Kamber, Micheline (2006). Data Mining: Concepts and Techniques. 2nd Edition: Morgan Kuaumann Publishers.

9. He, Zengyou, Deng, Shengchun, Xu, Xiaofei (2005). Improving K-Modes Algorithm Considering Frequencies of Attribute Values in Mode. Computational Intelligence and Security Lecture Notes in Computer Science. 3801, 157-162.

10. Huang, Zhexue (1997). Clustering large data sets with mixed numeric andcategorical values. Pacific-Asia Conference on Knowledge Discovery and Data Mining.

11. Khan, S., Shehroz, Ahmad, Amir (2013). Cluster Center Initialization for Categorical Data Using Multiple Attribute Clustering. Expert Systems with Applications. 40(18), 7444 7456.

12. Leela, V., Sakthipriya, K., Manikandan, R. (2013). A comparative analysis between k-mean and y-means Algorithms in Fisher s Iris data sets. International Journal of Engineering and Technology. 5(1), 245-249.

13. Liang, Jiye, Zhao, Xingwang, Li, Deyu, Cao, Fuyuan, Dang, Chuangyin. Determining the number of clusters using information entropy for mixed data. Pattern Recognition. 45, 2251-2265.

14. Liao, H., Ng, M., K. (2009). Categorical Data Clustering with Automatic Selection of Cluster Number. Fuzzy Information and Engineering. 1(1), 5 25.

15. Ng, K., Michael, Junjie, Mark, Huang, Zhexue, Joshua, Li, He, Zengyou (2007). On the Impact of Dissimilarity Measurein k-Modes Clustering Algorithm. IEEE Transactions on Pattern Analysis and Machine Intelligence. 29(3),503-507.

16. Pelleg, Dan, Moore, W., Andrew (2000). X-means: Extending K-means with Efficient Estimation of the Number of Clusters. Proceedings of the Seventeenth International Conference on Machine Learning. 727-734.

17. San, Mar, Ohn, Huynh, Van-Nam, Nakamori, Yoshiteru (2004). An Alternative Extension of the k-Means Algorithm for Clustering Categorical Data. International Journal Appl. Math. Computer. Science. 14(2),241 247.

18. Tibshirani, R., Walther, G., Hastie, T. (2000). Estimating the number of clusters in a dataset via the gap statistic. Technical Report 208, Department of Statistics, Stanford University, California.

19. Wagstaff, Kiris, Cardie, Claire, Rogers, Seth, Schroedl, Stefan (2001). Constrained K-means Clustering with Background Knowledge. Proceedings of the Eighteenth International Conference on Machine Learning. 577-584.

20. Xu, Rui, Wunsch, Donald (2005). Survey of Clustering Algorithms. IEEE Transactions on Neural Networks.16(3), 645-678.

21. Zhang, Chunfei, Fang, Zhiyi (2013). An Improved K-means Clustering Algorithm. Journal of Information & Computational Science, 10(1), 93 199.